Characterization of the mantle transcriptome in bivalves: Pecten maximus, Mytilus edulis and Crassostrea gigas

Characterization of the mantle transcriptome in bivalves: *Pecten maximus*, *Mytilus edulis* and *Crassostrea gigas*

Tejaswi Yarra[a,b,*], Karim Gharbi[a], Mark Blaxter[a], Lloyd S. Peck[b], Melody S. Clark[b]

[a] *University of Edinburgh, Institute of Evolutionary Biology, Ashworth Laboratories, Charlotte Auerbach Road, Edinburgh, EH9 3FL, UK*
[b] *British Antarctic Survey, Natural Environment Research Council, High cross, Magingley road, CB3 0ET, Cambridge, UK*

Abstract

The calcareous shells secreted by bivalve molluscs display diverse and species specific structural compositions, which indicates possible divergent biomineralization processes. Thus, studying multiple mollusc species will provide a more comprehensive understanding of shell formation. Here, the transcriptomes of the mantle tissues responsible for shell deposition were characterized in three commercially relevant bivalve species. Using high-throughput sequencing and bioinformatics tools, *de novo* transcriptome assemblies of mantle tissues were generated for the mussel *Mytilus edulis*, the oyster *Crassostrea gigas* and the scallop *Pecten maximus*. These transcriptomes were annotated, and contigs with similarity to proteins known to have shell formation roles in other species were identified. Comparison of the shell formation specific proteins in the three bivalves indicate the possibility of species specific shell proteins.

Keywords: molluscs, biomineralization, shells, shell formation

*Corresponding author
Email address: tejra@bas.ac.uk (Tejaswi Yarra)

1. Introduction

Bivalves are the second most speciose group in the phylum Mollusca [1] and are major components of marine food webs including wild harvested and farmed food sources. Bivalves are named after their distinctive twinned shells, which are usually attached at a hinge. The shells open and close with the help of ligaments and muscles, thus controlling the flow of water, nutrients and waste into or out of the organism [2]. The shell of bivalves plays an important role in supporting the living tissues of the animal and shielding it from the surrounding environment and predators [3].

The molluscan shell is a complex structure made of organic and mineral components. Generally, the outermost layer of a shell is the periostracum, which is made of organic material (such as conichilin) and makes up 1-5% of the shell weight. The inner layers are comprised of calcite and/or aragonite (polymorphs of calcium carbonate) and make up the rest of the 95-99% of shell weight [4]. The inner calcified layers are often structured in a variety of species specific orientations such as fibrillar prisms, cross lamellar, foliated [2], and in some bivalve species, the innermost layer is made up of hexagonal shaped aragonite nacre with a lustrous quality, valued as "mother of pearl". Importantly, these mineral layers include protein components that are responsible for nucleation of different calcium carbonate polymorphs and maintenance of shell integrity.

The majority of research on molluscan shell formation has utilized biochemical techniques. The protein component of the shell is separated from the mineral component and the dominant proteins are extracted and identified [5, 6]. Through biochemical and proteomic approaches, multiple proteins have been identified as being part of the shell matrix in molluscs. A few examples include nacerin, perlustrin, perculin [7], caspartin and calprismin [8], upsalin [9] and enzymes such as carbonic anhydrase [10]. Moreover, protein sequences were determined from techniques in proteomics and used to find corresponding genes, e.g. nacerin [11]. Currently, public sequence databases contain upwards of 500 identified proteins from the molluscan shell [12].

The molluscan shell is secreted by the mantle, a tissue that encloses the animal within the shell [1]. The mantle is a complex organ, comprising connective tissue, neural tissue and muscles in addition to glandular and epithelial components involved in shell secretion. Mantle transcriptomes, for the purpose of identifying biomineralization specific genes, have been assembled for clams [13], pearl oysters [14, 15, 16, 17], limpets [18], mussels [19] and scallops [20, 21, 22, 23]. The likely roles of some putative shell formation genes identified through mantle transcriptomes were explored through localization and developmental timing of expression. From some studies, there is indication that the structurally diverse molluscan shells are possibly underpinned by different genomic repertoires, as biomineralization pathways appear to have evolved independently several times in molluscs [24, 25]. Therefore, in order to understand shell deposition, biomineralization processes in multiple species should be investigated.

Three commercially relevant bivalve species are studied here: *Crassostrea gigas* (the Pacific oyster, family Ostreoida), *Mytilus edulis* (the blue mussel, family Mytiloida) and *Pecten maximus* (the king scallop, family Pectinoida). These bivalves are popular food sources and aquaculture of oysters, mussels and scallops accounted for 8.5 million tonnes of production worldwide in 2013 [26]. The three species have differing shell structures. The shell of *C. gigas* is made entirely of calcite (foliated calcite and calcite prisms) with a very thin periostracum that is usually lost *in vivo* [27]. The shell of *M. edulis* is comprised of aragonite nacre and fibrillar calcite with a thick periostracum that usually persists throughout life [2]. The shell of *P. maximus* contains an aragonite later in between two foliated calcite layers and a very thin periostracum that is usually lost *in vivo* [2].

In this paper, different parameters of transcriptome assembly are considered and the mantle transcriptomes are assembled for the three bivalves. The transcriptomes are then screened for potential biomineralization genes using sequence similarity searches against public databases to explore similarities and differences in putative biomineralization proteins between the three species.

2. Materials and methods

2.1. Sample collection

King scallops (*P. maximus*- length 92.90±2.82 mm, width 104.75±3.07 mm) were hand collected from Eilean Buidhe Island (Bull Loch, Scotland) and were maintained in indoor flow-through aquaria with a temperature of 15.22±0.47 $^{\circ}C$ and a salinity of 33.90±1.59 ppt. Blue mussels (*M. edulis*- length 48.73±4.16 mm, width 25.25±1.95 mm, height 19.82±2.14 mm) were harvested from wooden piling at Tarbert (East pier, Argyll, Scotland) and were also maintained in indoor flow through aquaria with a temperature of 15.40±0.43 $^{\circ}C$ and a salinity of 33.80±1.62 ppt. Pacific oysters (*C. gigas*- length 62.65±5.09 mm, weight 15.63±3.03 g) were manually harvested from lantern nets at Barmore Bay (Loch Fyne, Scotland) and were maintained in outdoor flow through aquaria at a temperature of 14.74±0.52 $^{\circ}C$ and a salinity of 36.00±0 ppt.

As part of a larger, ongoing shell repair experiment, three holes were drilled into each individual along the outer shell edge with even spacing using a cordless drill (Dremel, model 800, 1/8 inch for *P. maximus* and *C. gigas*, 1/32 inch for *M. edulis*) fitted with a round tipped end in order to not cut through to the tissue underneath. Shell repair was observed for four months and mantle tissue was collected (by destructive sampling of individuals) from both control and drilled individuals over a time course. For each sampled individual, the mantle edge was excised and snap frozen in liquid nitrogen and stored at -80 $^{\circ}C$. This report solely focusses on the characterization of the mantle transcriptomes of the three target species from both control and experimental samples collected at each of 13 or 14 time points during the course of the experiment (Supplementary Table 1). The larger shell repair program will be reported elsewhere.

2.2. RNA extraction and sequencing

Total RNA from the mantle tissues of *P. maximus* (n=14) was extracted using Tri-Reagent (Sigma-Aldrich) according to the manufacturer's instructions, and purified using RNeasy columns (Qiagen). Total RNA from the mantle tissues of *M. edulis* (n=14) and *C. gigas* (n=13) was extracted using SV Total

RNA isolation and purification kits (Promega). RNA samples were assessed for concentration and quality using a NanoDrop ND-100 Spectrometer (NanoDrop Technologies) and an Agilent 2200 Tapestation (Agilent Technologies). Library preparation and sequencing was carried out by Edinburgh Genomics (Edinburgh, UK). For each specimen sampled, RNA was converted to a sequencing library using the Illumina TruSeq stranded mRNA-seq library Prep kit (RNA input 1 µg, fragmentation time 8 min, 10 PCR cycles), and barcoded libraries were pooled and sequenced on an Illumina HiSeq 2500 (High output v4 mode) using 125 base paired-end reads, to generate from 10-20 million raw read pairs per sample.

2.3. Bioinformatics analysis

All analyses were carried out using default parameters unless otherwise specified. Adapters were trimmed from the reads using Trimmomatic v.0.33 [28]. The reads were further trimmed based on quality and length using Fastq-mcf v.1.04.636 [29] (setting the Phred quality score to 30 and minimum read length to 80 bp). The reads were normalized *in silico* with different coverage values and mantle contigs were assembled based on both the non-normalized and normalized reads using Trinity v.2.0.6 [30] (with SS_lib_type parameter set to RF to match the stranded library construction). Mantle contigs were assembled using the *de novo* mode for all three species, and additionally using the genome guided mode for *C. gigas* based on the published genome [31]. The read alignment bam file for input to the Trinity genome guided mode was generated using TopHat v.2.0.13 [32], and sorted using SAMtools v.1.1 [33].

Non-normalized raw reads were aligned to released mitochondrial sequences from RefSeq (19 JAN 2016), the different transcriptome assemblies created in this project, and the published *C. gigas* genome using TopHat to obtain the percentage of raw reads aligned to the assembled contigs. Using the Trinity pipeline, non-normalized raw reads were also aligned to the transcript assemblies using Bowtie v.1.1.1 [34] and abundance estimation was calculated using RSEM (RNA-Seq by Expectation-Maximization) v.1.2.20) [35].

Open reading frames of at least 100 codons, were identified in the contigs using Transdecoder (part of the Trinity pipeline). All protein similarity searches were carried out using BLAST (blastx or blastp) v.2.2.30 [36] with an E-value cutoff less than $1e^{-10}$ against SwissProt (10 JULY 2015) and Uniref90 (10 JULY 2015). The BLAST results were summarized based on the best similarity match for each transcript. Protein domains were identified using HMMER v.3.1b2 [37] and PFAM v.28.0 [38]. Signal peptides were identified using SignalP v.4.1 [39] and transmembrane regions identified using tmHMM v.2.0c [40]. The contigs and derived protein sequences were integrated using SQLite through Trinotate v.2.0 [41].

3. Results and discussion

3.1. Optimizing the assembly of mollusc mantle transcriptomes

RNA from 14 mantle samples from *Pecten maximus*, 13 mantle samples *Crassostrea gigas* and 14 mantle samples from *Mytilus edulis* were sequenced to yield 191 million, 217 million and 299 million read pairs respectively. After adapter trimming and quality and length filtering, 180 million, 208 million and 286 million read pairs remained for *P. maximus*, *C. gigas* and *M. edulis* respectively (Supplementary Table 1). The raw data are available in SRA under accession number SRP067223. Two approaches for mantle transcriptome assembly were explored: *In silico* read normalization and genome guided assembly.

The large volume of data produced by next-generation sequencing technologies can be very useful for providing depth in order to identify sequences expressed at low levels. However, analyses of these large datasets can be computationally challenging and random sub-sampling of the data to ease computation loads will lose information, as reads with low abundance may be lost. *In silico* normalization works by discarding highly abundant reads and is therefore preferable for reducing the amount of raw data without losing information of low abundance reads [41]. *In silico* normalized reads are only used for the purposes of transcript assembly and the complete set of non-normalized reads should be

used for downstream analysis such as differential expression or identification of single nucleotide polymorph sites.

Multiple versions of *C. gigas* transcriptomes were assembled to explore effects of *in silico* normalization on assembly metrics (Table 1). Normalization reduced the number of paired reads to be assembled by discarding all reads with abundance higher than the stated coverage values. 208 million paired reads were reduced to 16.9 million paired reads when normalized with a maximal coverage value of 30 fold, 28.8 million paired reads at 70 fold and 35.6 million paired reads at 100 fold. Trinity assembly of the non-normalized reads yielded more contigs than did assembly of normalized data, with increasing normalization stringencies reducing the number of contigs and increasing N50 lengths. While the assembly from the normalized data had fewer contigs, these contigs were on average longer, had greater N50 lengths, greater spans and comparable non-normalized cleaned read mapping rates (Table 1, Figure 1a). Similar analysis of the *P. maximus* and *M. edulis* transcriptome data showed the same trend of reduction in assembled contigs, but improved assembly metrics for normalized reads compared to non-normalized reads (Supplementary 2).

Instead of assembling a transcriptome *de novo*, assembled genome data can be used to condition the prediction of contigs given the prior expectation of the genome sequence. Reads mapped to a genome may also be used for downstream processes such as differential gene expression analysis or single nucleotide polymorphism identification. While high-quality genome assemblies are available for inbred model organisms, genome assemblies of non-model species can be compromised by heterozygosity and restricted access to resources. Therefore, the draft genome for *C. gigas* [31] was used to compare *de novo* and genome-guided assemblies.

Cleaned reads from *C. gigas* were mapped to the published genome to obtain an alignment information file for genome-guided assembly. Only 61.5% of the cleaned reads were mapped to the genome. Genome guided Trinity assemblies yielded fewer contigs with increased N50 lengths (Table 1, Figure 1b). Mapping the non-normalized clean reads to the genome-guided assemblies, revealed a

striking reduction in 10-15% of the raw reads map when compared to *de novo* transcriptome assemblies. The reduced read mapping rates may be caused by the high heterozygosity of the *C. gigas* individuals used in this study, incomplete or inaccurate scaffolding of the genome [42], or missing data in the genome. Less than 0.001% of the *C. gigas* raw reads mapped to mitochondrial sequences not from the genus *Crassostrea* and therefore there is very little contamination of the raw reads that could influence the raw read mapping rates to the assemblies.

The large number of contigs created by Trinity is not an unexpected behaviour and the number of contigs assembles can be influenced by multiple factors. The individuals used in this study were not inbred, and therefore a high amount of heterozygosity and polymorphism is expected, which in turn influences the number of contigs assembled. The high contig numbers can also be attributed to the Trinity algorithm capturing alternatively spliced transcripts derived from the same locus that differ in primary sequence. Partially spliced pre-mRNAs (containing unspliced introns) can also be captured and be reported as distinct transcripts, and losing contigs with such information may compromise biological interpretation.

3.2. Annotation of mollusc mantle transcriptomes

The *de novo* assemblies generated from reads normalized to 30 fold coverage were selected for deeper exploration of biomineralization in the three species, as these assemblies are smaller than non-normalized assemblies without excessive loss of read information. The assembled contigs are available through MolluscDB [43]. As the transcriptome assembly used individuals from different treatment groups, the transcripts were not categorized based on expression levels as the data would be influenced by the different treatment conditions. Instead, transcripts supported by low expression values were discarded and only transcripts with expression values above 1 FPKM (Fragments Per Kilobase of transcript per Million mapped reads) were considered for annotation (Table 2).

As expected, a higher proportion of *C. gigas* transcripts (42%) were found to have significant similarity to protein sequences in public databases than did *P.*

maximus (18%) and *M. edulis* (18%) due to the availability of a draft genome for *C. gigas*. Transcripts with open reading frames (ORFs) of at least 100 codons yielded higher sequence similarity based annotation levels at 95% for *C. gigas*, 70% for *M. edulis* and 80% for *P. maximus*. A large proportion of the annotation was derived from matches to *C. gigas* (Figure 2) and especially to the sequences published by the oyster genome project [31]. However, these annotation levels are probably a result of limited bivalve and mollusc information in public databases and it should not be considered that the three species have similar gene repertoires based solely on such sequence similarity results. Moreover, most of the oyster sequences in public databases were part of automated pipelines where proteins are only annotated based on known domains and therefore putative functional annotation is difficult to assign.

Comparison of Gene Ontology (GO) terms attributed to annotated protein sequences showed similar patterns across the three species (Supplementary Figure 1). The predicted functions were quite diverse as expected since the mantle is a functionally diverse organ. Some common functionalities of the mantle transcriptomes include regulatory and transcription factor sequences, nucleotide binding domains such as zinc fingers, muscle tissue related proteins such as myosin and actin, and many calcium-binding proteins that play roles in cell signalling [44]. The annotation reports of the transcripts and predicted ORFs, are included in the Supplementary information (Supplementary Tables 3, 4 and 5).

3.3. Putative biomineralization genes in mollusc mantle transcriptomes

Biomineralization relevant protein sequences were identified by screening the public databases for proteins observed in shell components or mantle tissue from studies focussed on molluscan shell formation. The mantle transcriptomes were screened for sequences similar to these biomineralization proteins (Table 3, Supplementary Table 6). If the protein predcited from a transcript contained a predicted transmembrane domain and the UniProt protein also had a transmembrane domain, these transcripts were indicated as membrane proteins. If

only the transcript-derived protein or the UniProt protein had a transmembrane domain, the transcripts was indicated as a putative membrane protein. Secreted proteins were identified in a similar manner (a protein was considered secreted if it contained a signal peptide domain and no transmembrane domains). This catalogue of putative biomineralization proteins revealed, in part, the similarities and differences in biomineralization proteins across these bivalves.

Comparing similarity between the three study species indicates that some biomineralization proteins appear to be species or genus specific. Only *C. gigas* sequences had significant sequence similarity to the Silk-like protein, Shelk2 and Nacrein-like proteins from *C. gigas* in the public database, while only *M. edulis* sequences has strong sequence similarity to proteins and domains such as mytilin, perwaplin and fibronectin, previously identified in *Mytilus* species. The sequence similarity results rarely identified 100% identity between the transcript from this study and the database representative from the same species. This could be due to alternate splicing, partial mis-assembly of contigs or natural population variance. In addition, paralogues of some genes like Nacrein may be differentially expressed by tissue and lifecycle stage. *P. maximus* transcripts from this assembly matched poorly to Nacrein proteins, with the strongest sequence similarity matches to Nacrein-like proteins from *P. vulgata*. This match could be explained because the *P. vulgata* Nacrein-like proteins were also identified through *de novo* transcriptome assembly of sequenced reads from a similar sequencing technology. This discrepancy of Nacrein proteins for *P. maximus* indicates that there are biases based on different sequencing technologies and sequences currently available in public databases.

4. Conclusions

We have generated three new transcriptome datasets relevant to the study of biomineralization in molluscs. We explored the outcomes of different assembly approaches, *in silico* normalization and genome guided assembly, and selected those derived from *in silico* normalization as being most effective. Genome

guided assemblies based on the current iteration of the published *C. gigas* genome were, surprisingly, not obviously superior to the *de novo* assemblies. Overall, annotation rates were similar across the three species, and reflected the diverse cell types and functions present in the complex mantle tissue. We identified potential species specific biomineralization proteins, but there are almost certainly novel genes that remain to be identified. These will be investigated in more detail in a time-course damage repair experiments which are currently under way in our laboratory.

Acknowledgements

Funding was received from Marie Curie Innovative Training Networks (ITN) (Grant agreement 605051). We would like to thank Gordon Goldsworthy from Loch Fyne Sea Farms for animal collection, Vicky Sleight from BAS for helping with experiment setup, and Kim Last, Christine Beveridge and Kati Michalek from SAMS for aquaria setup and animal husbandry. Sequencing was performed by Edinburgh Genomics, and we thank the laboratory and bioinformatics teams for support.

References

[1] E. Gosling, Bivalve Molluscs: Biology, Ecology and Culture, Fishing news books, Blackwell publishing.

[2] R. Bieler, P. M. Mikkelsen, T. M. Collins, E. A. Glover, V. L. González, D. L. Graf, E. M. Harper, J. M. Healy, G. Y. Kawauchi, P. P. Sharma, S. Staubach, E. E. Strong, J. D. Taylor, I. Tëmkin, J. D. Zardus, S. Clark, A. Guzmán, E. McIntyre, P. Sharp, G. Giribet, Investigating the bivalve tree of life – an exemplar-based approach combining molecular and novel morphological characters, Invertebrate Systematics 28 (2014) 32–15.

[3] F. Marin, G. Luquet, Molluscan shell proteins, Comptes Rendus Palevol 3 (67) (2004) 469 – 492. doi:http://dx.doi.org/10.1016/j.crpv.2004.07.009.

[4] J. D. Currey, The design of mineralised hard tissues for their mechanical functions, J. Exp. Biol. 202 (Pt 23) (1999) 3285–3294.

[5] B. Marie, A. Marie, D. J. Jackson, L. Dubost, B. M. Degnan, C. Milet, F. Marin, Proteomic analysis of the organic matrix of the abalone Haliotis asinina calcified shell, Proteome Sci 8 (2010) 54.

[6] J. C. Marxen, M. Nimtz, W. Becker, K. Mann, The major soluble 19.6 kDa protein of the organic shell matrix of the freshwater snail Biomphalaria glabrata is an N-glycosylated dermatopontin, Biochim. Biophys. Acta 1650 (1-2) (2003) 92–98.

[7] K. Mann, I. M. Weiss, S. Andre, H. J. Gabius, M. Fritz, The amino-acid sequence of the abalone (Haliotis laevigata) nacre protein perlucin. Detection of a functional C-type lectin domain with galactose/mannose specificity, Eur. J. Biochem. 267 (16) (2000) 5257–5264.

[8] F. Marin, R. Amons, N. Guichard, M. Stigter, A. Hecker, G. Luquet, P. Layrolle, G. Alcaraz, C. Riondet, P. Westbroek, Caspartin and cal-

prismin, two proteins of the shell calcitic prisms of the Mediterranean fan mussel Pinna nobilis, J. Biol. Chem. 280 (40) (2005) 33895–33908.

[9] P. Ramos-Silva, S. Benhamada, N. Le Roy, B. Marie, N. Guichard, I. Zanella-Cleon, L. Plasseraud, M. Corneillat, G. Alcaraz, J. Kaandorp, F. Marin, Novel molluskan biomineralization proteins retrieved from proteomics: a case study with Upsalin, Chembiochem 13 (7) (2012) 1067–1078.

[10] H. Miyamoto, T. Miyashita, M. Okushima, S. Nakano, T. Morita, A. Matsushiro, A carbonic anhydrase from the nacreous layer in oyster pearls, Proc. Natl. Acad. Sci. U.S.A. 93 (18) (1996) 9657–9660.

[11] H. Miyamoto, F. Miyoshi, J. Kohno, The carbonic anhydrase domain protein nacrein is expressed in the epithelial cells of the mantle and acts as a negative regulator in calcification in the mollusc Pinctada fucata, Zool. Sci. 22 (3) (2005) 311–315.

[12] N. C. for Biotechnology Information, Ncbi search (online query keywords: Mollusca, shell, protein,), National Center for Biotechnology Information. URL `http://www.ncbi.nlm.nih.gov/`

[13] M. S. Clark, M. A. Thorne, F. A. Vieira, J. C. Cardoso, D. M. Power, L. S. Peck, Insights into shell deposition in the Antarctic bivalve Laternula elliptica: gene discovery in the mantle transcriptome using 454 pyrosequencing, BMC Genomics 11 (2010) 362.

[14] Y. Shi, C. Yu, Z. Gu, X. Zhan, Y. Wang, A. Wang, Characterization of the pearl oyster (Pinctada martensii) mantle transcriptome unravels biomineralization genes, Mar. Biotechnol. 15 (2) (2013) 175–187.

[15] C. Joubert, D. Piquemal, B. Marie, L. Manchon, F. Pierrat, I. Zanella-Cleon, N. Cochennec-Laureau, Y. Gueguen, C. Montagnani, Transcriptome and proteome analysis of Pinctada margaritifera calcifying mantle and shell: focus on biomineralization, BMC Genomics 11 (2010) 613.

[16] Y. Deng, Q. Lei, Q. Tian, S. Xie, X. Du, J. Li, L. Wang, Y. Xiong, De novo assembly, gene annotation, and simple sequence repeat marker development using Illumina paired-end transcriptome sequences in the pearl oyster Pinctada maxima, Biosci. Biotechnol. Biochem. 78 (10) (2014) 1685–1692.

[17] S. Kinoshita, N. Wang, H. Inoue, K. Maeyama, K. Okamoto, K. Nagai, H. Kondo, I. Hirono, S. Asakawa, S. Watabe, Deep sequencing of ESTs from nacreous and prismatic layer producing tissues and a screen for novel shell formation-related genes in the pearl oyster, PLoS ONE 6 (6) (2011) e21238.

[18] G. D. Werner, P. Gemmell, S. Grosser, R. Hamer, S. M. Shimeld, Analysis of a deep transcriptome from the mantle tissue of Patella vulgata Linnaeus (Mollusca: Gastropoda: Patellidae) reveals candidate biomineralising genes, Mar. Biotechnol. 15 (2) (2013) 230–243.

[19] A. Freer, S. Bridgett, J. Jiang, M. Cusack, Biomineral proteins from Mytilus edulis mantle tissue transcriptome, Mar. Biotechnol. 16 (1) (2014) 34–45.

[20] M. Shi, Y. Lin, G. Xu, L. Xie, X. Hu, Z. Bao, R. Zhang, Characterization of the Zhikong scallop (Chlamys farreri) mantle transcriptome and identification of biomineralization-related genes, Mar. Biotechnol. 15 (6) (2013) 706–715.

[21] J. Ding, L. Zhao, Y. Chang, W. Zhao, Z. Du, Z. Hao, Transcriptome sequencing and characterization of Japanese scallop Patinopecten yessoensis from different shell color lines, PLoS ONE 10 (2) (2015) e0116406.

[22] S. Artigaud, M. A. Thorne, J. Richard, R. Lavaud, F. Jean, J. Flye-Sainte-Marie, L. S. Peck, V. Pichereau, M. S. Clark, Deep sequencing of the mantle transcriptome of the great scallop Pecten maximus, Mar Genomics 15 (2014) 3–4.

[23] X. Sun, A. Yang, B. Wu, L. Zhou, Z. Liu, Characterization of the mantle transcriptome of yesso scallop (Patinopecten yessoensis): identification of genes potentially involved in biomineralization and pigmentation, PLoS ONE 10 (4) (2015) e0122967.

[24] D. J. Jackson, C. McDougall, K. Green, F. Simpson, G. Worheide, B. M. Degnan, A rapidly evolving secretome builds and patterns a sea shell, BMC Biol. 4 (2006) 40.

[25] D. J. Jackson, C. McDougall, B. Woodcroft, P. Moase, R. A. Rose, M. Kube, R. Reinhardt, D. S. Rokhsar, C. Montagnani, C. Joubert, D. Piquemal, B. M. Degnan, Parallel evolution of nacre building gene sets in molluscs, Mol. Biol. Evol. 27 (3) (2010) 591–608.

[26] Fisheries and aquaculture department, Global production statistics (online query), Food and Agriculture Organization of the United Nations.
URL `http://www.fao.org/fishery/statistics/global-production/query/en`

[27] Y. Dauphin, A. D. Ball, H. Castillo-Michel, C. Chevallard, J. P. Cuif, B. Farre, S. Pouvreau, M. Salome, In situ distribution and characterization of the organic content of the oyster shell Crassostrea gigas (Mollusca, Bivalvia), Micron 44 (2013) 373–383.

[28] A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (15) (2014) 2114–2120.

[29] E. Aronesty, Command-line tools for processing biological sequencing data, ea-utils.
URL `http://code.google.com/p/ea-utils`

[30] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren,

C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol. 29 (7) (2011) 644–652.

[31] G. Zhang, X. Fang, X. Guo, L. Li, R. Luo, F. Xu, P. Yang, L. Zhang, X. Wang, H. Qi, Z. Xiong, H. Que, Y. Xie, P. W. Holland, J. Paps, Y. Zhu, F. Wu, Y. Chen, J. Wang, C. Peng, J. Meng, L. Yang, J. Liu, B. Wen, N. Zhang, Z. Huang, Q. Zhu, Y. Feng, A. Mount, D. Hedgecock, Z. Xu, Y. Liu, T. Domazet-Lo?o, Y. Du, X. Sun, S. Zhang, B. Liu, P. Cheng, X. Jiang, J. Li, D. Fan, W. Wang, W. Fu, T. Wang, B. Wang, J. Zhang, Z. Peng, Y. Li, N. Li, J. Wang, M. Chen, Y. He, F. Tan, X. Song, Q. Zheng, R. Huang, H. Yang, X. Du, L. Chen, M. Yang, P. M. Gaffney, S. Wang, L. Luo, Z. She, Y. Ming, W. Huang, S. Zhang, B. Huang, Y. Zhang, T. Qu, P. Ni, G. Miao, J. Wang, Q. Wang, C. E. Steinberg, H. Wang, N. Li, L. Qian, G. Zhang, Y. Li, H. Yang, X. Liu, J. Wang, Y. Yin, J. Wang, The oyster genome reveals stress adaptation and complexity of shell formation, Nature 490 (7418) (2012) 49–54.

[32] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S. L. Salzberg, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, Genome Biol. 14 (4) (2013) R36.

[33] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools, Bioinformatics 25 (16) (2009) 2078–2079.

[34] B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (3) (2009) R25.

[35] B. Li, C. N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinformatics 12 (2011) 323.

[36] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410.

[37] R. D. Finn, J. Clements, W. Arndt, B. L. Miller, T. J. Wheeler, F. Schreiber, A. Bateman, S. R. Eddy, HMMER web server: 2015 update, Nucleic Acids Res. 43 (W1) (2015) W30–38.

[38] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman, R. D. Finn, The pfam protein families database (Jan 2012).

[39] T. N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: discriminating signal peptides from transmembrane regions, Nat. Methods 8 (10) (2011) 785–786.

[40] A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J. Mol. Biol. 305 (3) (2001) 567–580.

[41] B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. Leduc, N. Friedman, A. Regev, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, Nat Protoc 8 (8) (2013) 1494–1512.

[42] D. Hedgecock, G. Shin, A. Y. Gracey, D. Van Den Berg, M. P. Samanta, Second-Generation Linkage Maps for the Pacific Oyster Crassostrea gigas Reveal Errors in Assembly of Genome Scaffolds, G3 (Bethesda).

[43] L. Stevens, M. Blaxter, Molluscdb: A transcriptome database for mollusc species (2016).
URL `http://molluscdb.afterparty.bio.ed.ac.uk/`

[44] W. J. Chazin, Relating form and function of EF-hand calcium binding proteins, Acc. Chem. Res. 44 (3) (2011) 171–179.

TABLES AND FIGURES

	de novo assemblies				Genome-guided assemblies	
	Non-normalized	normalized100	normalized70	normalized30	Non-normalized	normalized30
Cleaned reads:						
Total reads (paired, million)	208.1	35.6	28.8	16.9	208.1	16.9
Total bases (paired, billion)	24.8	4.3	3.5	2.0	24.8	2.0
Total Trinity transcripts	664,459	630,469	619,842	577,369	251,855	200,024
Total Trinity genes	476,096	428,784	410,018	350,250	223,799	157,945
%GC	37.54	37.68	37.71	37.77	38.51	39.17
Statistics of Trinity transcripts:						
N50	557	623	655	773	796	1,094
Min length (bp)	224	224	224	224	224	224
Max length (bp)	27,871	30,497	38,423	32,940	19,856	26,783
Median length (bp)	327	334	338	359	354	411
Total assembled bases (Mbp)	336	337	339	344	152	144
Mapping of cleaned (non-normalized) reads:						
Overall alignment %	84.8	83.5	83.7	80.1	71.4	69.6

Table 1: Trinity assemblies of *C. gigas* with different parameters

Note 1: normalized(n) = the cleaned reads were normalized with the max_coverage parameter set to n

	P. maximus	*C. gigas* (*de novo*)	*M. edulis*
Trinity assembly from reads normalized at 30x coverage After removal of lowly expressed transcripts			
Total Trinity transcripts	228,088	426,028	559,818
Total Trinity genes	150,241	280,305	399,890
%GC	36.96	38.14	33.51
Statistics of Trinity transcripts:			
N50	1,083	814	566
Min length (bp)	224	224	224
Max length (bp)	17,211	32,940	25,276
Total assembled bases (Mbp)	154	253	282
Sequence similarity search:			
SwissProt	27,053	65,515	60,495
UNIREF90	41,254	177,294	98,383
Protein sequences (ORF >100aa):	40,715	112, 066	121,472
Sequence similarity search:			
SwissProt	22,794	51,435	45,628
UNIREF90	30,478	100,433	69,414
Pfam domains	32,518	87,083	92,194
Signal peptide domains:	2,430	8,867	7,105
Transmembrane domains:	6,501	19,739	17,371

Table 2: Mantle transcriptome assembly and annotation

Note 1: All e-values for sequence similarity searches against public databases were set to $1e^{-10}$

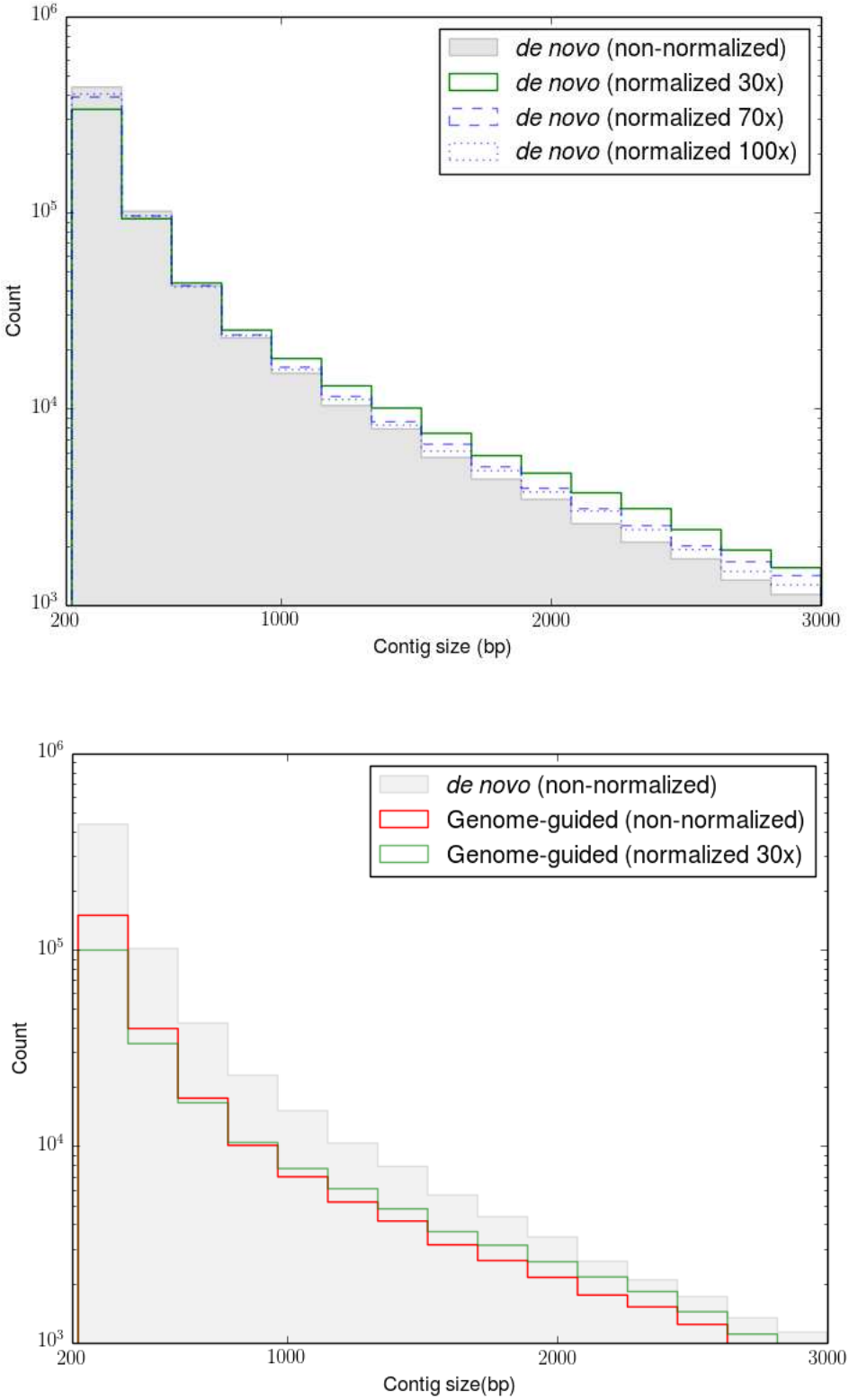

Figure 1: Contig lengths (for contigs up to 3000 bases long) vs. count (log-scale) of different *C. gigas* Trinity assemblies. (a) Non-normalized vs. normalized: Fewer contigs of smaller length were yielded from normalized assemblies. The normalized assemblies also have more contigs of longer lengths compared to the non-nomralized assembly. (b) *de novo* vs. genome guided: Overall fewer contigs of all lenghts were produced by the genome-guided assembly. The genome-guided assembly of normalized reads had fewer contigs of smaller length compared to genome-guided assembly of non-normalized reads.

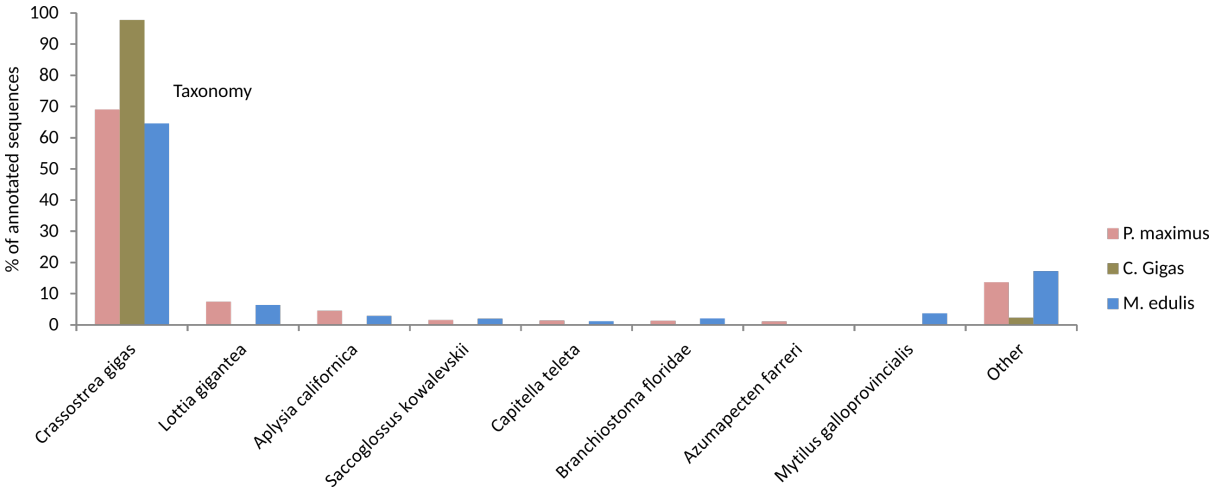

Figure 2: Annotation categorized by Taxonomy

			Percent identity		
Name	**UniProtID(s)**	**S/M**	***P. maximus***	***C. gigas***	***M. edulis***
Pinctada fucata (Pearl oyster)					
Amorphous calcium carbonate binding	A6XBS1		26	30	33
Basic protein N23	I1Z9K2	pS			53**
Calcium-transporting ATPase	B2KKR0, B2KKR1, B2KKR2	M	85	91	88
Calmodulin-like protein	Q3BDI8		67	70	71
Chitin synthase	A7BIC0	pM	76*	69**	68**
Nacre proteins	L8B5V5, G1K3T5, G1K3U4	pS			35
Nacrein-like proteins	A0ZSF2, Q27908	S		34**	34**
Protein PIF	C7G0B5	pS	27**	27**	28**
Tyrosinase-like	A7BK18, A1IHF0, A1IHF1		33	37	34
Pinctada margaritifera (Black-lipped pearl oyster)					
Asparagine-rich protein	H2A0M0	pS	26**	27**	28**
Asparate-rich protein	H2A0M1	S		49**	
BPTI/Kunitz domain-containing	H2A0M2, H2A0N1, H2A0N9, H2A0P0, H2A0N5	pS	39*	44**	40*
EGF-like domain containing	H2A0L2, H2A0L3	S	25	41	25
Fibronectin type III domain-containing	H2A0L7, H2A0L8	pS	39	47	44
Mantle protein	H2A0K7	S		38	33
Nacrein-like proteins	F1DS85, F5B6X1, F5B6X2, F5B6X3, F5B6X4	pS	28**	36**	38**
Peroxidase-like protein	H2A0M7	pS	39*	41	37*
Protein PIF	H2A0N4	pS	27**	27**	29**

			Percent identity		
Name	**UniProtID(s)**	**S/M**	***P. maximus***	***C. gigas***	***M. edulis***
Putative amine oxidase	H2A0M3	pS	43	49	43
Putative beta-hexosaminidase	H2A0L6	pS	26	32*	32*
Putative chitinase	H2A0L4, H2A0L5	pS	44	55	45
Tyrosinase-like	H2A0L1, H2A0L0	pS	32	40	34
Valine-rich protein	H2A0K6	S		43	
Pinctada maxima (White-lipped pearl oyster)					
BPTI/Kunitz domain-containing	P86959, P86963, P86964	pS	59**	39*	44**
EGF like domain containing	P86954, P86953	S	27	43	26
Mantle protein	P86948	S		42	34
Nacrein-like protein	A0ZSF3	S		35*	36*
Putative beta-hexosaminidase	P86956	pS	27*	32*	33*
Putative chitinase	P86955	pS	45	56	46
Tyrosinase-like protein	P86952	pS	32	38	34
Valine-rich protein	P86947	S		43	
Lottia gigantea (Giant owl limpet)					
Calcium-transporting ATPase	V4AKV4	M	81	83	82
EGF-like domain-containing	B3A0R6, B3A0S3	pS	28	25	28**
Glycine, glutamate and proline-rich protein	B3A0P5	pS		46**	52**
Putative carbonic anhydrase	B3A0P2	S	28	26	23
Peroxidase-like proteins	B3A0Q8	pS	39	37	38
Putative PPIase	B3A0R0	pS	63	64	57
SCP domain-containing protein	B3A0P7, B3A0P8	S	29**	28**	33**
Haliotis discus hannai (Japanese abalone)					
Acetylcholine binding proteins	B3SNJ8, B5KGU8	pM	33	34	32

			Percent identity		
Name	**UniProtID(s)**	**S/M**	***P. maximus***	***C. gigas***	***M. edulis***
Haliotis laevigata (Abalone)					
Perlucin	P82596	pS	40	40	39
Haliotis asinina (Donkey's ear abalone)					
BPTI/Kunitz domain-containing	P86733	pS	53	52	45
Ependymin-related protein 1	P86734	pS	27		26
Crassostrea gigas (Pacific oyster)					
Calcium transporting ATPase	K1QA13	M	83	97	85
Chitin synthases	K1QG38, K1Q3B7, etc.	pM	68	82	71*
Gigasins	P86784, P86785, P86786, P86787, P86788, P86789		39*	99	38
Nacrein-like proteins	R9WGX8, K1RJ02, etc	pS	48**	98	48*
Shelk2 subtype 7	G9M4P0	pM		99	
Silk like protein	G9M4L4	pS		100	
Mytilus californianus (California mussel)					
Fibronectin type III domain-containing	P86861	pS	57*	49	98
Nacrein-like protein	P86856	pS	25	41	70
Mytilin	P86858, P86859	pS			78
Shell matrix protein	P86860	pS	38	35	97
Mytilus galloprovincialis (Mediterranean mussel)					
BMSP	G1UCX0		27**	27**	100**
Chitin synthase	A5HKN1, Q27W11	pM	73*	64**	99**
Lectin	B3EWR1		52		97
Mytilin	P86853	S			82
Perlucin-like	P86854	pS	29	35	84*
Perlwapin-like	P86855	pS			63

			Percent identity		
Name	**UniProtID(s)**	**S/M**	***P. maximus***	***C. gigas***	***M. edulis***
Mytilus edulis (Blue mussel)					
EP Protein	Q6UQ16	pS	28	28	96
Biomphalaria glabrata (Freshwater snail)					
Dermatopontin	P83553	pS	33	41	39
Ruditapes philippinarum (Japanese littleneck clam)					
Insoluble matrix shell proteins	P86987, P86982	pS	34	54	47
Patella vulgata (Common limpet)					
AP24 protein	J7Q5J5		36*		32*
Lustrin A (Fragment)	J7QAX0		44*	48**	41*
Nacrein-like proteins	J7QAX2, J7QJU2, J7QXI7	pS	49	51	51
Mizuhopecten yessoensis (Japanese scallop)					
Calcium-transporting ATPase	K1QA13, O96039	M	95	97	85
Nacrein-like protein	A0ZSF5, A0ZSF5	S		34**	37**
Hyriopsis cumingii (Triangle sail mussel)					
Alpha-2-macroglobulin	A0A1G5		50	48	50**
Apolipophorin	R4VDM5			43	43
Chitin deacetylase isoforms	J7FHX7, J7FHI0	PS	39**	58*	56**
Perlucin	M9QW24	PS	40	32	51

Table 3: Percent identity of bivalve mantle deduced proteins compared to shell proteins deposited in Uniprot.

Note 1: * indicates that the 60-80% of the protein is aligned to. ** indicates <60% of protein is aligned to. No mark indicates >80% of protein aligned to.

Note 2: (p)S – (putative) secretory proteins, (p)M – (putative) membrane proteins

SUPPLEMENTARY TABLES AND FIGURES

Sample name	*P. maximus*	*C. gigas*	*M. edulis*
Day 0 - Drilling of shells			
Sample 001	Drilled Day 1	Drilled Day 1	Drilled Day 1
Sample 002	Drilled Day 2	Drilled Day 2	Drilled Day 2
Sample 003	Drilled Day 3	Drilled Day 3	Drilled Day 3
Sample 004	Drilled Day 5	Drilled Day 5	Drilled Day 5
Sample 005	Drilled Day 7	Drilled Day 7	Drilled Day 7
Sample 006	Drilled Day 10	Drilled Day 10	Drilled Day 10
Sample 007	Drilled Day 14	Drilled Day 14	Drilled Day 14
Sample 008	Drilled Day 21	Drilled Day 29	Drilled Day 21
Sample 009	Drilled Day 29	Drilled Day 72	Drilled Day 29
Sample 010	Drilled Day 71	Control Day 1	Drilled Day 71
Sample 011	Drilled Day 111	Control Day 3	Drilled Day 111
Sample 012	Control Day 1	Control Day 7	Control Day 1
Sample 013	Control Day 10	Control Day 72	Control Day 28
Sample 014	Control Day 111	-	Control Day 111
Sequenced reads information (paired, million):			
Raw reads per sample	13.6±3.4	16.7±2.4	21.4±5.5
Total raw reads	191.0	217.3	299.1
Cleaned reads per sample	12.9±3.4	15.2±4.3	20.4±5.2
Total cleaned reads	180.4	208.1	286.0

Table S1: Sample description and sequenced read information

Note 1: Cleaned reads were trimmed of standard Illumina adapters, and filtered based on Phred quality score of 30 and a minimum length of 80 bases

	P. maximus		*M. edulis*	
	Non-normalized	normalized30	Non-normalized	normalized30
Cleaned reads:				
Total reads (paired, million)	180.4	9.6	285.9	20.6
Total bases (paired, billion)	21.7	1.2	34.5	2.5
Total Trinity transcripts	350,171	296,632	926,978	800,070
Total Trinity genes	252,206	178,457	671,740	509,809
%GC	36.85	37.17	33.38	33.51
Statistics of Trinity transcripts:				
N50	785	1151	487	604
Min length (bp)	224	224	224	224
Max length (bp)	16,392	17,211	24,652	25,276
Median length (bp)	337	388	318	343
Average length (bp)	581	709	466	526
Total assembled bases (billion)	0.204	0.210	0.432	0.421

Table S2: Non-normalized and normalized transcriptome assemblies for *P. maximus* and *M. edulis*

Note 1: normalized(n) = the cleaned reads were normalized with the max_coverage parameter set to n

Table S3,S4,S5 - Excel files: Trinotate reports of the annotation

Table S6 - Excel file: Sequence similarity of known shell/mantle proteins, BLAST result details and Contig/Protein sequence identifiers from *P. maximus* , *C. gigas* and *M. edulis* mantle transcriptome

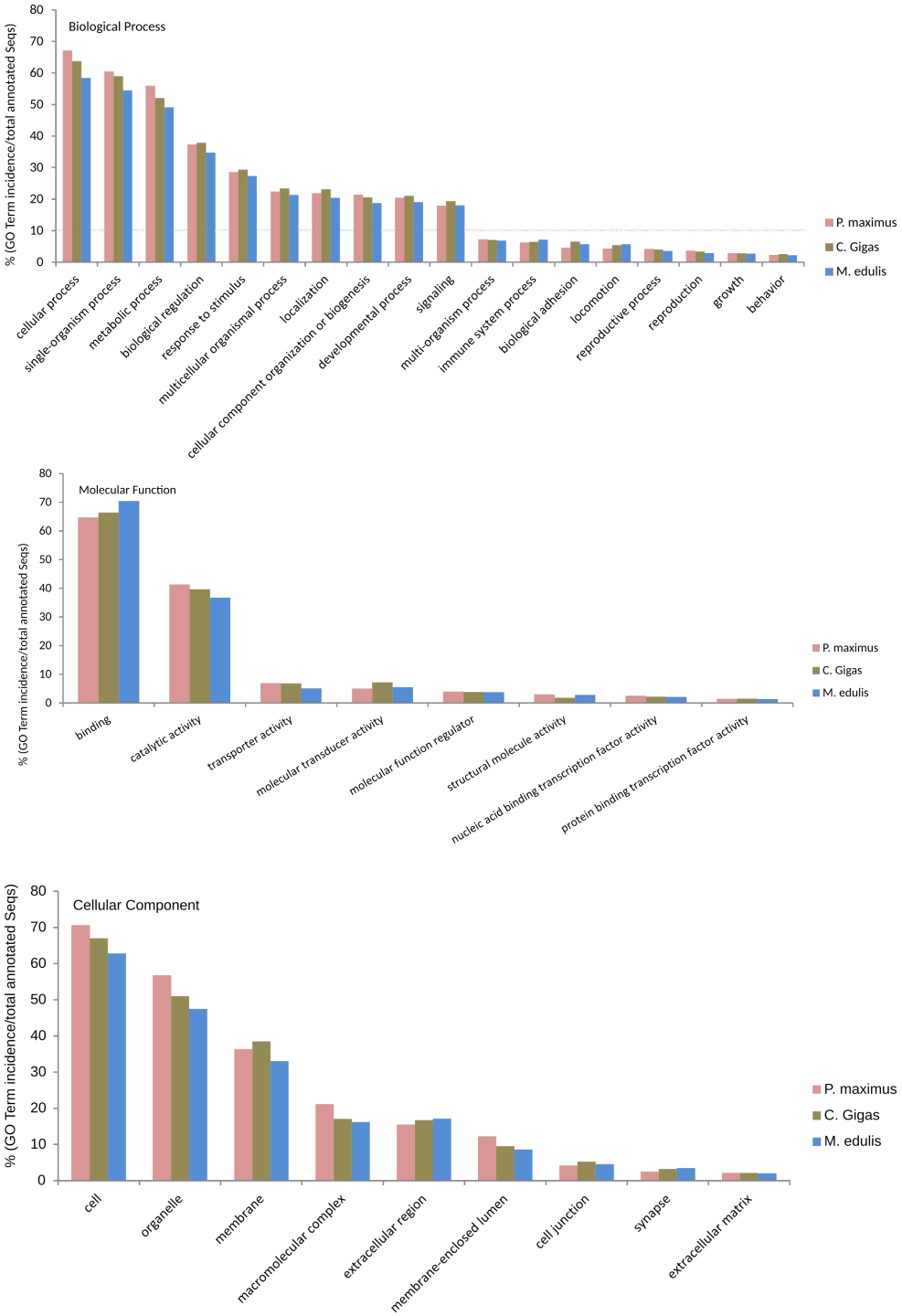

Figure S1: Annotation categorized by GO categories